Erschienen in: Butt, Miriam (Hrsg.): Proceedings of KONVENS 2006 (Konferenz zur Verarbeitung natürlicher Systeme), Universität Konstanz. - Konstanz: Universität Konstanz, 2006. S. 125-129.

Eine Ontologie für die Grammatik – Modellierung und Einsatzgebiete domänenspezifischer Wissensstrukturen

Roman SCHNEIDER
Institut für deutsche Sprache (IDS)
R5, 6-13
D-68161 Mannheim
schneider@ids-mannheim.de

Einleitung

Ontologien erfreuen sich seit Mitte der Neunziger Jahre nicht nur in der KI-Forschung, sondern insbesondere im Umfeld von Knowledge Management und Information Retrieval großer Popularität. Nicht allein die vergleichsweise junge Vision eines zukünftigen „Semantic Web", sondern gerade das bereits heute ausgeprägte Verlangen, Wissensstrukturen effizient zu verwalten und bei der Exploration großer Informationsmengen gezielt einzusetzen, haben zu dieser Popularität beigetragen. Ontologie-basierte Anwendungen unterstützen in vielerlei Kommunikationssituationen das Gelingen von Interaktion. Hierzu gehören der Informationsaustausch zwischen verschiedenen Software-Systemen, Schnittstellen zur Mensch-Maschine-Interaktion oder auch der Dialog zwischen menschlichen Kommunikationspartnern mit differierenden terminologischen Vokabularien.

Hierzu existieren mittlerweile eine Vielzahl domänenspezifischer Ontologien für einzelne Fachgebiete und Branchen. Der vorliegende Beitrag beleuchtet Fragestellungen hinsichtlich der Entwicklung und Anwendung einer solchen Domänen-Ontologie am Beispiel des grammatischen Informationssystems *grammis*.

Das Projekt *grammis* ist in der Abteilung Grammatik des Instituts für Deutsche Sprache (IDS) in Mannheim beheimatet und hat zum Ziel, ein umfassendes hypermediales Informationssystem zur deutschen Grammatik bereitzustellen, auf das weltweit über das Internet zugegriffen werden kann. Derzeit umfasst es vier Komponenten. Inhaltliches Kernstück ist die „Systematische Grammatik". In dieser Komponente wird versucht, ein Gesamtbild der Grammatik der deutschen Gegenwartssprache zu entwerfen. Die verwendeten Fachtermini erläutert ein „Terminologisches Wörterbuch"; das „Grammatische Wörterbuch" liefert Spezialinformationen zu einzelnen Lemmata. Abgerundet wird das System durch eine „Grammatische Bibliografie". Sämtliche Inhalte liegen in Form strukturierter Hyperdokumente vor und werden gemeinsam mit Meta-Informationen innerhalb eines objekt-relationalen Datenbankmanagementsystems verwaltet.[1]

Die *grammis*-Ontologie soll unter Verwendung texttechnologischer Methoden[2] eine Grundlage für die inhaltliche Textklassifikation und das Text-Retrieval im *grammis*-Fachkorpus schaffen. Aussagekräftige Zusammenhänge zwischen terminologischen Konzepten sollen explizit erfasst und auf dieser Basis auch implizit kodierte Fakten abgeleitet werden können.

1 Bestimmung der Relationstypen

Für die Entwicklung der *grammis*-Ontologie wurden folgende Relationstypen verwendet:[3]

Synonymie: Zwischen zwei oder mehreren Bestandteilen eines Konzepts besteht eine Synonymie-Beziehung, falls sich eine Austauschbarkeit in mindestens einem fachspezifischen Kontext konstatieren lässt.

Eigenschaft: Die Eigenschaftsbeziehung ist in der *grammis*-Ontologie eine binäre Relation zwischen zwei Konzepten. Konzepte besitzen

[1] Weitere ausführliche Informationen zur Konzeption und Implementierung von *grammis* bietet Schneider (2004).

[2] Ein umfassender Überblick über Methoden und Anwendungen der Texttechnologie findet sich in Lobin/Lemnitzer (2004).

[3] Vgl. hierzu auch Schneider (2006).

demzufolge (obligatorische oder fakultative) Eigenschaften, die selbst im Konzeptinventar enthalten sind – aber nicht notwendigerweise in eine Hierarchie eingebunden sein müssen – und von allgemeineren auf speziellere Konzepte vererbt werden.

Hyponymie: Die transitive Über- bzw. Unterordnung von Konzepten stellt die wichtigste Möglichkeit der Hierarchiebildung dar. Wesentlich für die *grammis*-Ontologie sind polyhierarchische Strukturen, d.h. die Zulassung mehrerer übergeordneter Konzepte zu einem Hyponym. Dabei muss allerdings das Auftreten logischer Widersprüche – etwa in Form von Schleifen – vermieden werden, d.h. ein neues Hyponym darf nicht bereits als direktes oder indirektes Hyperonym eingetragen sein. Mit Hilfe eines „baumartigen Netzes" gerichteter Graphen kann dann später etwa eine Klassifikation von Fachtexten erfolgen: Werden innerhalb eines Textes mehrere Konzepte mit einem gemeinsamen Hyperonym gefunden, bietet sich in vielen Fällen dieses Hyperonym als Kandidat für die Textcharakterisierung an.

Meronymie: Eine Teil-Ganzes-Beziehung besteht dann, wenn ein Konzept konstituierender Bestandteil eines anderen Konzepts ist. Aufgrund der Transitivität des Beziehungstyps können auf diese Weise mehrstufige Hierarchien aufgebaut werden. Analog zur Eigenschaftsbeziehung lässt sich zwischen obligatorischer und fakultativer Meronymie unterscheiden: So kann man beispielsweise davon ausgehen, dass jede Nominalphrase zwingend ein Nomen, aber nicht notwendigerweise ein attributives Adjektiv oder einen Artikel beinhalten muss.

Autorenschaft: Dieser Relationstyp dient der Verknüpfung von Konzepten, die wissenschaftliche Problemstellungen, Modelle oder Teildisziplinen ausdrücken, mit einschlägigen Autoren. Maßgeblicher Beweggrund für die Modellierung dieser Relation war die Absicht, einflussreiche Persönlichkeiten, linguistische Fachautoren und Fachliteratur aus elektronischen Bibliografien konzeptabhängig zu erschließen.

Instanziierung: Instanz-Beziehungen bestehen beinahe ausschließlich zwischen Wortart-Konzepten und einzelnen Wortformen aus dem „Grammatischen Wörterbuch" von *grammis*.

Assoziation: Für diesen Relationstyp bestehen keine inhaltlichen Vorbedingungen. Er wird dann eingesetzt, wenn zwei Konzepte de facto miteinander in Verbindung stehen, die Art dieser Verbindung aber nicht präzise formulierbar oder aus quantitativer Perspektive vernachlässigbar erscheint.

2 Bestimmung der Konzepte

Für die Erstellung der *grammis*-Ontologiebasis wurde erfolgreich eine kombinierte Methode eingesetzt, die statistische Auswertungen, linguistische Analysetechniken sowie eine manuelle Nachbearbeitung umfasst. Das zugrunde liegende Fachkorpus besteht aus den XML-strukturierten Hypertexteinheiten des grammatischen Informationssystems *grammis* sowie der ebenfalls am IDS publizierten Propädeutischen Grammatik *ProGr@mm*. Insgesamt handelt es sich dabei um ca. 2.000 miteinander verknüpfte Hypertexteinheiten mit knapp 1.000.000 Wortformen. Weiterhin kam das Korpus-Recherchesystem COSMAS zum Einsatz, mit dessen Hilfe am IDS derzeit ca. 160 allgemeinsprachliche Korpora mit über 1,6 Milliarden laufenden Wortformen verwaltet werden. Nachfolgend die aufeinander aufbauenden Schritte zur Konzeptbestimmung:

1) Frequenzanalyse Fachkorpus: Als Input dienen die Hypertexteinheiten des Fachkorpus (FK). Für alle laufenden Wortformen, unter Auslassung der in einer Ausnahmeliste hinterlegten Stoppwörter (z.B. Konnektoren wie *und*, *aber*, *also* etc.), wird die Frequenz bestimmt. Wortformen mit einem Frequenzwert, der unterhalb eines vorher festgelegten Schwellenwerts liegt, werden getilgt. Der Output besteht aus einer geordneten, zweispaltigen Liste (Wortform und FK-Frequenz).

2) Markup-Analyse: Als Input dienen die Wortformenliste aus Schritt 1 sowie die XML-kodierten[4] Meta-Informationen der Fachkorpus-Hypertexteinheiten. Wortformen, die in prominenten Textstrukturen erscheinen, also z.B. in Titeln, Zwischenüberschriften, Definitionsabsätzen oder semantisch etikettierten Hyperlinks, erhalten einen „Bonus" und rücken in der Rang-

[4] Eine Beschreibung der Markup-Sprache *grammisML* findet sich in Schneider (2004, S. 251ff).

liste auf. Der Output besteht aus einer entsprechend modifizierten Wortformenliste.

3) Frequenzanalyse allgemeinsprachlicher Korpus: Als Input dient die Wortformenliste aus Schritt 2 sowie die in COSMAS verwalteten allgemeinsprachlichen Texte. Zu jeder Wortform in der Liste wird der Frequenzwert im allgemeinsprachlichen Korpus (AK) bestimmt. Der Output besteht aus einer dreispaltigen Liste (Wortform, FK-Frequenz, AK-Frequenz).

4) Auffälligkeitsanalyse: Als Input dient die Wortformenliste aus Schritt 3. Unter Zuhilfenahme eines erprobten Algorithmus[5] wird für jede Wortform ein „Weirdness"-Wert nach folgender Formel errechnet:

$$\tau(w) = N_{AK} \, f_{FK} \; / \; f_{AK} \, N_{FK}$$

N_{AK} = Gesamtzahl Wortformen im AK

N_{FK} = Gesamtzahl Wortformen im FK

f_{AK} = Wortform-Frequenz im AK

f_{FK} = Wortform-Frequenz im FK

Der errechnete Wert liefert eine Aussage darüber, welche Wortformen signifikant häufiger im Fachtext-Korpus als im allgemeinsprachlichen Korpus auftreten. Je höher der Wert, desto wahrscheinlich handelt es sich um einen domänenspezifischen Konzeptkandidaten. Der Output besteht aus einer Liste der Konzeptkandidaten, wobei auch wieder diejenigen Wortformen aussortiert werden, die unterhalb eines festgelegten Schwellenwerts liegen.

5) Kollokationsanalyse: Als Input dienen die Kandidatenliste aus Schritt 4 sowie das Fachtextkorpus. Mit Hilfe einer Kollokationsanalyse lässt sich darin das gemeinsame Auftreten von Konzeptkandidaten untersuchen. Da die Hypertexteinheiten durchgehend und einheitlich strukturiert sind, können bei diesem Schritt variable Umgebungsgrenzen (Sätze, Absätze, Hypertexteinheiten) gewählt werden. Prinzipiell lassen sich sogar bereits erste gerichtete Relationen konstatieren: Tritt ein Konzeptkandidat X signifikant häufiger gemeinsam mit einem Konzeptkandidaten Y auf als Kandidat Y mit X, so liegt die Vermutung nahe, dass Y für ein allgemeineres Konzept und X für ein spezielleres Konzept steht. Der Output dieses Schritts liefert Konzeptkandidaten-Cluster, d.h. zu jedem der Kandidaten eine Menge von Kollokationspartnern.

6) Relationsbestimmung: Als Input dient die Clusterliste aus Schritt 5. Ein menschlicher Experte kann nun bewerten, welche der Konzeptkandidaten für die Beschreibung der Domäne in Betracht kommen und welche Relationen auf Grund der Konzeptcluster modelliert werden sollen. Als Ergebnis erhält er ein vorläufiges Konzeptnetz, das bereits einige Teilhierarchien beinhaltet.

3 Anwendungsszenarien

Die Ideen und Vorschläge hinter der populären Vision vom „Semantic Web" rücken einen Umstand in die öffentliche Diskussion, der seit Jahren zu den offenkundigen Defiziten der derzeitigen Internet-Struktur zählt: Natürlichsprachlich kodierte Angebote auf Webseiten lassen sich zwar aus der modernen Medienrealität nicht mehr wegdenken und bieten dem menschlichen Anwender einen im Vergleich mit traditionellen Publikationsmethoden deutlich schnelleren Zugang zu aktuellen Informationen aus zahlreichen Fachgebieten. Doch nicht nur wer persönlich via Internet-Browser Suchmaschinen wie Google zur Informationsrecherche benutzt, wird zunehmend von der Menge und Unübersichtlichkeit des Angebots erdrückt. Auch programmierte Spezialisten – Agents, Crawlers, Robots oder Spiders genannt – müssen immer häufiger vor dem Unterfangen kapitulieren, formal und inhaltlich heterogene Webinhalte zuverlässig zu klassifizieren und automatisch diejenigen Angebote auszuwählen, die am besten zu einem bestimmten Informationsbedürfnis passen.

An diesem Punkt kommen nun elektronisch verfügbare Ontologien ins Spiel. Die *grammis*-Ontologie wird derzeit auf dreierlei Weise dazu eingesetzt, Probleme der Mensch-Maschine-Kommunikation hinsichtlich der Informationserschließung anzugehen sowie „intelligenter"[6] Software den Zugang zu domänenspezifischen Inhalten zu erleichtern:

[5] Vgl. die ausführliche Beschreibung in Gillam/Tariq/Ahmad (2005).

[6] Im Sinne von: „kombinations- und lernfähig"; vgl. hierzu auch Schneider (2004, S. 17 ff).

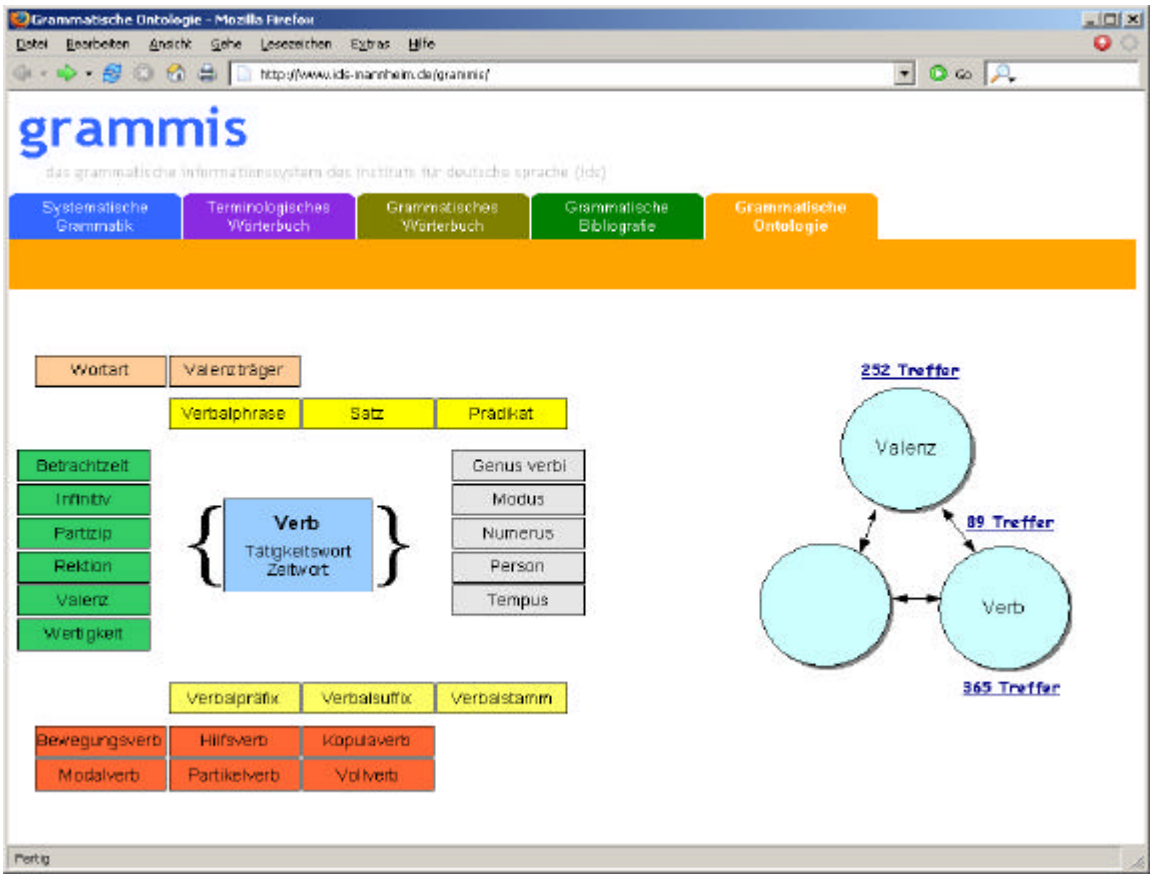
Inhaltliche Klassifizierung: Wie bereits einführend dargestellt, wurde das grammatische Informationssystem *grammis* ursprünglich mit der Zielsetzung konzipiert, grammatische Fachinhalte in einer für menschliche Benutzer verständlichen Form erschließbar zu machen. Analysen der Zugriffs-Logfiles zeigen jedoch, dass zunehmend auch programmierte Robots und Agentenprogramme versuchen, auf diese Angebote zuzugreifen. Die Möglichkeit solcher Programme, textuelle Inhalte auszuwerten und ihrem Auftraggeber verlässliche Ergebnisse zu liefern, hängt wesentlich von der Verwendung eines gemeinsamen Terminologie-Vokabulars ab. Um die Kommunikation zwischen *grammis* und externer Software zu erleichtern, wird mit Hilfe der *grammis*-Ontologie automatisch ein Schlagwort-basiertes Abstract jeder *grammis*-Informationseinheit generiert. Hierzu werden sämtliche in der Einheit gefundenen Konzept-Termini sowie die in der Ontologiebasis hinterlegten Relationen zwischen diesen Konzepten ausgewertet. Das Abstract enthält dann primär diejenigen Konzepte, die in der Relationenhierarchie als „Top-level terms" markiert sind.

Hilfestellung beim Text-Retrieval: Auch die Volltextsuche, für menschliche Benutzer häufig die bevorzugte Einstiegsvariante in ein komplexes Informationssystem, liefert nur dann befriedigende Resultate, wenn Mensch und Computer die gleiche Sprache sprechen, d.h. eine gemeinsame Terminologie verwenden. Gibt der Anwender als Suchwort „Ergänzung" ein, sollte *grammis* erkennen, dass es sich hierbei um ein Synonym für „Komplement" handeln kann und eine Verbindung zum Konzept „Valenz" herstellen. Der Suchbegriff wird also zunächst expandiert und die Menge der Treffer dadurch vergrößert. Um zu vermeiden, dass durch Verallgemeinerung die Anzahl der selektierten Hypertexteinheiten unverhältnismäßig zunimmt, ist ab einem bestimmten Level die umgekehrte Strategie notwendig: Erkennt das System, dass der Suchbegriff in der Relationenhierarchie weit oben steht und gleichzeitig übermäßig viele Treffer generiert wurden, kann es – ggf. nach Rückfrage beim Benutzer – eine Auswahl untergeordneter bzw. verwandter Konzepte zur Eingrenzung der Suchanfrage vorschlagen. Ausgehend von „Valenz" würde sich also beispielsweise der Schritt zu Hyponymen wie „Verbvalenz" oder anderweitig verknüpften Konzepten wie „Reduktionstest" anbieten.

Bereitstellung eines grafischen Recherche-Frontends: Die grafische Repräsentation einer Ontologiebasis[7] unterstützt den menschlichen Anwender durchgehend in allen Phasen des Ontologie-Lebenszyklusses. Sowohl die Erstellung bzw. Modifizierung von Konzepten wie auch die spätere Exploration und Anwendung profitieren aus leicht nachvollziehbaren Gründen von einer adäquaten Visualisierung der Inhalte. Insbesondere in Benutzungssituationen, in denen der Benutzer sein Informationsbedürfnis nicht präzise formulieren kann oder einfach einmal ungezielt „stöbern" möchte, befördert ein strukturierter grafischer Überblick die Orientierung. Für das Informationssystem *grammis* wurde deshalb ein spezielles Frontend geschaffen, das die Konzepte und Relationen der Ontologiebasis variabel darstellt und das Navigieren und Recherchieren innerhalb des *grammis*-Informationsraums ermöglicht.

Die nachfolgende Darstellung illustriert die Funktionsweise: Im Zentrum der Übersicht steht das aktuell angesteuerte Konzept, gemeinsam mit den Synonym-Begriffen. Oberhalb finden sich, in farblich abgestuften Blockelementen, die unmittelbaren Hyperonyme und Holonyme; darunter versammeln sich Hyponyme und Meronyme. Eigenschaftsbeziehungen werden seitlich versetzt dargestellt, ebenso wie unspezifisch assoziierte Konzepte. Per Mausklick können nun die verschiedenen Relationen aktiviert und der Standort im Informationsraum verändert werden. Für die eigentliche Recherche stehen die drei kreisförmigen Container auf der rechten Seite bereit. Per Drag-and-Drop lassen sich beliebige Konzeptelemente in die Container ziehen. Das System recherchiert daraufhin direkt in der *grammis*-Hypertextbasis sowie in der ebenfalls angeschlossenen Bibliografie. Unmittelbar neben den Containern erscheint die Anzahl der Treffer für das einzelne Konzeptelement als Hyperlink, an den Kanten zwischen den Containern die ebenfalls via Hyperlink weiterverfolg-

[7] Prominente Verfahren beruhen z.B. auf *fish-eye-views* oder *hyperbolic trees*. Einen Überblick über gängige Methoden der Visualisierung von Ontologien bieten z.B. Fluit/Sabou/van Harmelen (2003).



baren Ergebnisse der kombinierten Recherche nach mehreren Konzeptbegriffen.

Die Leistungsfähigkeit der beschriebenen Anwendungsschnittstellen hängt naturgemäß nicht allein von Umfang und Güte der zugrunde liegenden Ontologie ab, sondern auch von den übrigen sprachtechnologischen Komponenten. Ontologie-basiertes Text-Retrieval und die Text-Klassifikation erfordern beispielsweise einen angemessenen Umgang mit Flexionsformen und Komposita. Die *grammis*-Recherche sowie die anderen *grammis*-spezifischen Textanalyse- und Klassifikationsmodule verwenden deshalb einen integrierten Mechanismus zur Lemmatisierung und Stammformenerweiterung. Für die Zukunft sind darüber hinaus vielfältige Erweiterungen denkbar. Ein interessantes Entwicklungspotenzial verspricht die Berücksichtigung sprachübergreifender Anwendungsgebiete. International orientierte Forschungsarbeit profitiert von multilingual konzipierten Terminologiedatenbanken, zu denen mehrsprachige Ontologien einen wesentlichen Beitrag leisten können.

Literatur

Fluit, C. / Sabou, M. / van Harmelen, F. (2003): Supporting User Tasks through Visualis ation of Lightweight Ontologies. In: Staab, S. / Studer, R. (Hg.), Handbook on Ontologies in Information Systems. Berlin: Springer.

Gillam, L. / Tariq, M. / Ahmad, K. (2005): Terminology and the construction of ontology. In: Terminology. Volume 11, Number 1, S. 55-81.

Lobin, H. / Lemnitzer, L. (2004): Texttechnologie. Perspektiven und Anwendungen. Tübingen: St auffenburg.

Schneider, R. (2004): Benutzeradaptive Systeme im Internet: hformieren und Lernen mit GRAMMIS und ProGr@mm. Mannheim: IDS (=amades 4/04).

Schneider, R. (2006): Texttechnologie und Grammatik. In: Breindl, E. / Gunkel, L. / Strecker, B. (Hg.): Grammatische Untersuchungen, Analysen und Reflexionen. Tübingen: Narr.